‘Phantom’ compositional effects in English school value-added measures

Perry, Thomas

DOI:
10.1080/02671522.2018.1424926

Document Version
Peer reviewed version

Citation for published version (Harvard):
Perry, T 2019, ''Phantom' compositional effects in English school value-added measures: the consequences of random baseline measurement error', *Research Papers in Education*, vol. 34, no. 2, pp. 239–262. https://doi.org/10.1080/02671522.2018.1424926

Accepted Manuscript

‘Phantom’ Compositional Effects in English School Value-Added Measures: The Consequences of Random Baseline Measurement Error

Thomas Perry

School of Education,
University of Birmingham,
Birmingham
B15 2TT
UK
Email: t.w.perry@bham.ac.uk

A compositional effect is when pupil attainment is associated with the characteristics of their peers, over and above their own individual characteristics. Pupils at academically selective schools, for example, tend to out-perform similar-ability pupils who are educated with mixed-ability peers. Previous methodological studies however have shown that ‘phantom’ compositional effects can arise purely from measurement error. Through simulations using English National Pupil Database data, this paper demonstrates that moderate rates of measurement error in baseline scores can produce a systemic bias in value-added (‘Progress’) scores that favours higher ability pupils. This bias is large enough to explain the ostensibly superior performance of selective ‘grammar’ schools that is seen in the English data and create an appreciable bias across the school performance distribution. This paper describes how such biases can arise from ‘regression attenuation’ and discusses the more general methodological implications of the results. The paper concludes that the consequences of baseline measurement error on school value-added scores is greatly mitigated by controls for school-level prior attainment, but that the English ‘Progress’ value-added measures, as they do not control for prior attainment at school-level, are seriously biased and that the observed ‘grammar school effect’ is likely to be largely or wholly spurious.

Key Words: measurement error; attenuation bias, compositional effects, Progress 8, school value-added; education policy

1. Introduction

Measurement error in prior attainment measures has been found to give rise to a statistical problem known as attenuation (or dilution) bias which can give rise to so-called 'phantom' compositional effects. The central aim of this paper is to estimate the consequences of this potential problem for the English accountability system. To this end, this paper presents a simulation of the effect of pupil-level random measurement error on school-level value-added scores using English National Pupil Database (NPD) data. The results strongly suggest that the impact of attenuation bias is sufficient to explain compositional effects apparent in English data, most clearly manifested in a longstanding yet variable advantage enjoyed by English selective 'grammar' schools in terms of their value-added scores.

Before proceeding, it is important to note that measurement error is not necessarily nor even presumably random (Gorard 2011a), nor will it be confined to baseline attainment measures. Moreover, this paper's focus on random baseline measurement error is not to downplay the importance of biases related to other contextual factors (e.g. pupil disadvantage), which recent research shows to be substantial (Perry 2016). The present focus on random error in baseline attainment measures is due to the remarkable finding of this paper that realistic rates of error in baseline scores, even while making benign assumptions about error randomness, can produce sizable and systemic biases in school value-added scores. The specific problem of attenuation bias and random baseline measurement error, therefore, warrants especial attention. This paper concludes that attenuation bias is both a serious and fundamental problem for the English measures and an important methodological issue for value-added estimation more generally. In overview, this paper:

1. Reviews literature on measurement error in value-added measures. This starts by briefly discussing the seriousness and nature of error in general. It then considers previous studies which have examined the effect of school composition and the difficulties involved in measuring it. These studies have implicated attenuation bias as a potential problem, and so a short explanation of attenuation bias is given.
2. Describes the practical context of the English accountability system (2004-2016) by presenting estimates of compositional effects in official English Key Stage 2 (age 11) to Key Stage 4 (age 16) (KS2-4) value-added measures. These results are valuable in their own right and also highlight the impacts of relevant policy moves during this period.

3. Presents of simulation study using English NPD data examining the impact of random measurement error on school-level value-added scores. The main results are followed by subsidiary analyses examining the extent to which it is possible to correct for the problem.
4. The paper closes by considering the implications for the use of value-added methods in general and for the English accountability system in particular.

2. Bias, Error and Phantom Compositional Effects in Value-Added Measures

2.1 The Nature and Seriousness of Error

School value-added scores capture *unexplained* variation in student outcomes, left over after controlling for selected non-school factors such as prior attainment. It is only by assumption that this residual largely or wholly reflects the relative performance of schools. Bias and error can have particularly serious consequences for value-added estimates because they introduce extraneous and generally unobserved sources of variation into this residual variation. Issues of missing data and error are well documented, particularly in relation to measures of pupil characteristics used in 'contextualised' value-added measures (Gorard 2010).

Even simple value-added approaches such as that used in the new English 'Progress 8' measure (see DfE 2016) require a valid measure of prior attainment and outcome attainment, in this case at KS2 and KS4 respectively. Yet, within any given testing system there is a considerable degree of unreliability stemming from marking and other factors that are not relevant to the pupils' performances (Newton 2013; He, Hayes, and Wiliam 2013). Any error within the underlying measures of attainment will also introduce error into value-added estimates.

The central question, therefore, is how serious the consequences of error and bias are for any given value-added estimates. Some authors have reached the conclusion that error and bias renders value-added methods of little value and even harmful (Gorard 2006b; Gorard 2010). As well as detailing numerous sources of bias and error, Gorard (2010) observes that value-added scores are created by finding a difference between actual and expected performance, both of which are measured with error, the former directly (in an examination) and the latter because of the errors in the measured variables (e.g. prior attainment) which are used to form the attainment expectation. Both of these errors can be either positive or negative and so can 'propagate' as well as cancel out. Moreover, the difference between expected and actual performance (with expected value zero) will be considerably smaller than the original exam scale. Propagation combined with the smaller magnitude of the new value-added scale

will mean that the error on value-added estimates can be much larger relative to error on the original examination scores, which Gorard argues will already be sizable.

This negative position on error is contested, and led to a series of exchanges between Gorard and prominent educational effectiveness researchers (Gorard 2010; Muijs et al. 2011; Gorard 2011a; Gorard 2011b; Reynolds et al. 2012). Defenders of value-added, over the course of these exchanges with Gorard, propounded their view that value-added school effects are sufficiently large, stable and consistent to be of value and that the error tends to be a much smaller component of the variance that, moreover, tends to be random (Reynolds et al. 2012) and amenable to technical solutions (Muijs et al. 2011). These exchanges raised several issues around bias and error within value-added estimation. The present paper pursues one such issue, that of the impact of random measurement error on school value-added scores, introduced below.

2.2 The Impact of Random Error

One major source of the differing evaluations of the magnitude of the problem of error in the methodological debates (above) is a fundamental disagreement over whether it is true, as Reynolds et al. (2012, 116) claim, that errors 'tend to be randomly distributed'. If this is the case, the argument goes, pupil-level errors will generally cancel out and error is 'unlikely to be systematically different in different schools'.

Reynolds et al. (2012) go on to cite several studies that have examined the influence of random measurement error within multi-level models (Woodhouse et al. 1996; Goldstein, Kounali, and Robinson 2008; Ferrão and Goldstein 2009). These studies give a consistent picture of the effects of random measurement error in multi-level value-added models, suggesting that when models are adjusted for a lower reliability (see Ferrão and Goldstein 2009, 954) of the measure of prior attainment (ranging from 1 down to 0.6 across studies) this leads to a) an increase in the prior attainment coefficient values, b) a decrease in pupil-level variance, c) an increase in the intra-school correlation, an indicator of the importance of the school effect, and d) school-level variance being largely unaffected. As Reynolds et al. (2012) point out, these studies suggest that the size of the relative school effect (at least as measured by the intra-school correlation) is increased when adjusting for measurement error and, citing Goldstein, Kounali and Robinson (2008, 256), that the effects of error are 'confined to the fixed part of the model' (i.e. confined to the factor coefficient estimates and not affecting the residual, or 'random', school effects). Similarly, Ferrão and Goldstein (2009) conclude that measurement error influences parameter estimates (e.g. the prior attainment coefficient) but not the value

added estimates. They find high correlations of 0.97 or higher between school value-added estimates with and without considering measurement error.

These points are discussed later in this paper in light of the empirical results. At present, it is worth reiterating that while such studies, including the present study, provide an important test of the likely effects of error, this is a somewhat limited one because errors are not necessarily (or even presumably) random (Gorard 2011a, 18). As Ferrão and Goldstein (2009, 963) also note, their study does not address variation in measurement across the score scale nor 'unobserved patterns of measurement error variation across schools'. Therefore, studies of random measurement error provide a fairly optimistic scenario, considering only a single threat to validity under what are quite benign assumptions.

2.3 Phantom Compositional Effects

The simulation presented later in the paper finds that the introduction of random errors in prior attainment scores can lead to systematic positive biases relating to pupil attainment. It also suggests that outside of a simulation these biases are only visible in analysis at a higher level: in this case, school-level (i.e. for school averages). This finding links the present study with a strand of research examining so-called 'phantom' compositional effects, addressed here.

A compositional effect is when pupils' attainment is found to be associated with the characteristics of their peers, over and above the association on an individual basis. Research has examined whether pupils perform better than otherwise-comparable pupils when educated with peers who are more able or who have higher socio-economic status (see for examples, Marks 2015; Timmermans and Thomas 2014; Boonen et al. 2014). There are several plausible reasons why compositional effects may have an impact of attainment (see Harker and Tymms 2004), but the evidence is currently not conclusive, despite claims to the contrary from authors such as Liu et al. (2015, 423) for whom compositional effects have been 'consistently verified'. Rather, as Nash (2003, 444) noted back in 2003, 'it can seem that for every analysis that finds a composition effect there is another that does not'. This situation remains unchanged, with school compositional effects still proving 'controversial' (Reynolds et al. 2014, 209).

Numerous major studies over many years have failed to find appreciable effects (Lavy, Silva, and Weinhardt 2012; Boonen et al. 2014; Marks 2015). Gibbons and Telhaj (2012, 26), for example, conclude that 'group composition matters little' and that the negligibly small estimates found in their particular study were 'in line with the effect sizes in most other studies on the topic' (see Gibbons and Telhaj 2012, 33, Table 1). Similarly, recent studies of

compositional effects related to peer socio-economic status (SES) have concluded 'that school-SES effects are trivial and do not warrant a policy response' (Marks 2015, 18).

A major methodological issue within this area of research is that of measurement error. As Gorard (2006a) notes, the difficulties of estimating school effects using noisy data are already sizable and these are compounded when studying compositional effects which are a fraction of the overall school effect. There are now several methodological studies that have shown that spurious (or 'phantom') compositional effects can arise from pupil-level measurement error (Harker and Tymms 2004; Hutchison 2007; Pokropek 2014; Televantou et al. 2015) or the failure to take pupil-level control variables such as SES into account (Nash 2003).

Various explanations have been given for the mechanism that produces or inflates compositional effects. Harker and Tymms (2004, 195) explain the effect in terms of the 'unreliability of the predictor variables':

> "As the pupil-level measure becomes less reliable, it explains less of the variance. But the aggregate measure does not loose (sic) its reliability to anything like the same extent, since it is an average across a group of pupils. The aggregate measure is then well placed to mop up unexplained variance at the school level."
>
> (Harker and Tymms 2004, 192)

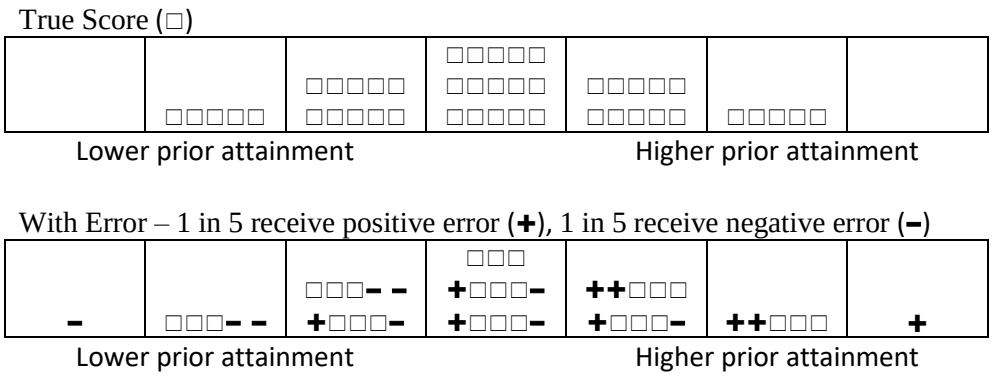
Hutchison (2007) presents a mathematical proof showing the impact of measurement error on estimates at pupil and school-level. Crucially, he demonstrates that phantom compositional effects can arise even in the absence of any omitted variable biases and even as sample sizes approach infinity. The latter means that approaches to quantify (actual or hypothetical) sampling error such as confidence intervals or corrections such as Bayesian shrinkage (see Snijders and Bosker 2011, 62) will not solve the problem. Moreover, this means that *any amount* of measurement unreliability will exert downward bias on the coefficient values of the prior attainment measure (see Hutchison 2007, 224, point 1) and 'reliably give rise to spurious [compositional effects]' (230).

2.4 The Problem of Attenuation Bias

This downward bias on the coefficient values has been referred to elsewhere (e.g. van Ewijk and Sleegers 2010, 249) as 'attenuation bias' or 'dilution bias' (Frost and Thompson 2000) because measurement error dilutes the predictive power of, in the present case, the prior attainment variable when estimating future attainment. The function of the prior attainment

variable is to discriminate between pupils of different abilities to produce statistical expectations of future performance for given values of prior attainment. The presence of measurement error will mean that for any observed prior attainment score there will be three types of pupil: first, pupils whose prior attainment score error was approximately zero and for whom their observed score is their 'true' score. Second, 'under-achievers' whose error was negative and whose true prior attainment score is therefore higher than their observed score. Third, 'over-achievers' whose error was positive with a true score lower than their observed score. One might presume that the proportions of over- and under-achievers will tend to balance and have no systematic effect on the mean final attainment at a given point on the prior attainment scale. However, when the score in question is above the highest point (mode) on a normally distributed prior attainment score distribution, there will be more pupils immediately below than above; this means that the over-achievers, with a lower *true* score, tend to be more numerous. To illustrate this, Figure 1 (below) shows a simple distribution of true scores followed by the same distribution for which 1 in every 5 pupils receives a negative error (**–**) and gets a lower score and 1 in every 5 pupils receives a positive error (**+**) and receives a score higher than their true score. Note how the varying frequencies across the normal distribution result in over-achievers outnumbering under-achievers above the modal score and vice versa below the modal score. Also note that in practice over- and under-achievers will be unobserved.

Figure 1 – Random Error Across Normally Distributed Scores



Relating this to the value-added ('Progress') approach, if one were to then estimate the mean final attainment score at a point above the modal score in the prior attainment scale, the mean final attainment score for these pupils will be attenuated (or diluted) by the greater proportion of pupils with a lower *true* prior attainment score. Below the mean score, the opposite applies: any given point will have more pupils with a higher true prior attainment score.

This problem means that measurement error in the prior attainment variable will shrink the mean final attainment scores for pupils at all points on prior attainment scale towards the overall final attainment mean. This will, in turn, inflate the absolute value (positive or negative) of all value-added scores derived from comparing actual final attainment scores to this expectation. At the extreme, should the prior attainment score be unavailable or completely dwarfed by measurement error and therefore contain no information, it is impossible to distinguish higher or lower ability pupils. In this scenario, the best expectation of final attainment will be the overall mean and value-added scores will simple be the mean-centred final attainment scores. The crucial theoretical point here is that, as measurement error increases, one *tends towards* this extreme position; an imperfect prior attainment score will result in an expectation only partially adjusted for ability rather than an unbiased but 'noisy' expectation. The question, addressed here, is whether the effect will have a serious impact in the given context of the English school value-added scores. Further explanation and illustration of attenuation bias is presented in connection with the results of the measurement error simulation in this paper.

There are examples of studies picking up the problem of attenuation bias in other contexts and other value-added methods (e.g. Dieterle et al. [2015] in the context teacher value-added and McCaffrey, Castellano, and Lockwood [2015], when using student growth percentile approaches). More recent studies of phantom compositional effects have also highlighted this attenuation problem. Televantou et al. (2015, 80) point out that 'models that fail to adjust for measurement error in the individual-level variable lead to a negative bias in its estimated coefficient at the individual level' but note that direction of the bias at school-level can be either positive or negative. Another study in the area, Pokropek (2014), further discusses the theory behind phantom compositional effects, again highlighting the issue of attenuation bias. As Pokropek (2014, 7) notes, the magnitude of the bias will depend on the specific measures and context but that 'overestimation of contextual effect might be expected in a majority of scenarios'. The present study examines its seriousness and consequences in the specific context of the English school accountability system. The starting point is looking at compositional effects in official English value-added performance data from 2004 to 2016.

3. Compositional Effects in English Value-Added Measures 2004-2016

The following results use school-level performance data from 2004 to 2016 for all maintained, mainstream schools in England. To examine possible compositional effects, Table 1, below, presents school-level correlations between English value-added scores and 1) the KS2 measure

used as a baseline for the VA measure and 2) the measure of final attainment used as the outcome. The correlation between VA and final attainment estimates the extent to which VA gives unique information about school performance (Gorard 2006b). Some level of correlation is expected with final attainment because schools with higher value-added will also tend to get higher final attainment. Any correlation between VA and prior attainment, however, suggests either a (real or phantom) compositional effect.

Table 1 - Correlation between KS2-4 value-added measures, prior attainment and final attainment 2004-2016[ǂ] (School-level Data)

Year	KS2-KS4 Value-added Measure	KS2 Average Point score	Capped GCSE and equivalents point score.	Selective School Effect[†]
2004	KS2- Equivalents VA score	.50	.85	24.8
2005	KS2 to KS4 contextual value added score	.01	.52	6.5
2006		.03	.43	3.3
2007	KS2 to KS4 contextual value added score with shrinkage factor.	.00	.42	2.8
2008		.01	.29	3.4
2009		.01	.27	4.0
2010		.00	.27	5.9
2011	Best 8 VA measure	.17	.75	16.8
2012		.13	.76	15.9
2013		.15	.77	17.5
2014		.29	.75	23.7
2015		.38	.79	25.6
2016[§]	Progress 8	.31	.74	0.35 (≈ 21 on former scale)*

[ǂ] See Appendix A, Table A2 for details of specific NPD data and variables used

[†] Mean Value-added for schools with mean KS2 average point scores of 31 and above

[§] Using Attainment 8 as the KS4 measure.

* Multiplied by 60 (10*6) due to changes in score scales (see DfE 2017)

There are several points to note from Table 1: First, there are particularly high correlations in 2004. This was found in previous research by Gorard (2006b, 239) who found a 'surprising correlation' between the school VA scores and the school-level final attainment scores in the 2004 data. Analysis of the 2004 pupil- and school-level data during this study found that the *pupil-level* correlation between (KS2-4) value-added and KS2 prior attainment was zero (i.e. the model successfully corrected for prior attainment), whereas the school-level correlation for

the same variables was 0.50. This is a clear example of the problem presently discussed and the important theoretical implication of the present results that correcting for prior attainment at pupil-level is insufficient to produce unbiased school value-added estimates.

Second, these results show that since the contextualised value-added measure (CVA) (see Evans 2008) was discontinued and school-level associations between performance and value-added were no longer taken into account, English value-added scores have not been entirely independent of intake prior attainment. The CVA models included a school-level mean KS2 score and its standard deviation (Kelly and Downey 2010) so would necessarily remove any observable compositional effects.

Third, the compositional effect has grown in the last two years along with the selective school effect. Compositional effects were lower in 2011-2013 compared to 2014, 2015 and, to a lesser extent, 2016. This change is most likely to have come from a fall in the predictive value of the corresponding KS2 scores. Such a fall in predictive value will stem from either measurement error at KS2 or the alignment of the KS2 and KS4 tests, where reforms to 2009-11 KS2 tests or 2014-16 KS4 tests reduced the predictive value of the former for the latter. It is very unlikely that a compositional effect has arisen in English system in little over two years including grammar schools collectively increasing their performance by an average of around one and a half GCSE grades per pupil per school. It is far more believable to conclude that the reliability of the KS2 scores or their ability to predict KS4 scores has deteriorated. It is also interesting to note the change in the correlation between VA and final attainment between 2007 and 2008. Again, changes in measurement are a far more likely explanation than wholesale, systematic changes in the performance of English schools.

4. Simulation of the Effect of Random Measurement in Baseline

4.1 Simulation Aims

The literature reviewed above raises question of whether random pupil-level measurement errors will, rather than cancelling out at school-level (Reynolds et al. 2012), produce systematic biases in school-level results due to their magnitude or due to attenuation bias. The extent to which pupil-level measurement error will have serious or systematic effects will depend on the specific context (Pokropek 2014) of the reliability of measures, school sizes and the distribution of pupils to schools (i.e. the composition of schools) and whether errors are in fact random, as claimed. Estimating the seriousness of these problem in the specific context of the English data is therefore of great value. It is increasingly recognised that the valid use (Kane 2013) of value-

added measures in any given evaluation system requires clear information about the quality of specific measures and their likely rates of error (AERA 2015). This raises an applied research question and the focus for the present study.

4.2 Research Question

What is the effect of random pupil-level measurement error in Key Stage 2 (KS2) prior attainment scores on English school-level Key Stage 2 to 4 (KS2-4) value-added scores?

4.3 Data Source and Measures: Pupil-Level English National Pupil Database 2013

The simulation presented below uses an extract from the 2013 National Pupil Database (NPD). The NPD is a large dataset containing performance and pupil background data for all state-educated pupils in England, about 93% of the school population, over many years. A large number of variables are collected relating to achievement and to pupil and school characteristics. This study uses only a school unique reference number and the following three variables (see DfE 2013, for further details):

- *Key Stage 4 (KS4) 'Best 8' (plus bonus) score:* The official measure of pupil performance at age 16. This is a composite measure combining scores across pupils' best 8 General Certificate of Secondary Education (GCSE) examination scores, with the Mathematics and English scores doubled to give these greater weight.
- *Key Stage 2 (KS2) average point score:* The official measure of pupil performance at age 11. For most English pupils this score is recorded in the final year of primary school before attending a separate secondary school. KS2 scores are therefore used as the baseline measure for secondary value-added measures. The KS2 APS scores used are an average of pupils' English, Maths and Science point scores; although note that current KS2 has changed and in future will include only mathematics and reading scores. As these are from the 2013 NPD extract, KS2 APS still uses the pre-2010 measure (see RAISEonline, 2014).
- *'Best 8' value-added score:* The official pupil-level value-added score estimating value-added between Key Stage 2 and 4. This is used solely for calibration and comparison purposes.

The analysis was conducted for state-maintained, mainstream schools. Schools with KS4 cohorts of less than 15 pupils were dropped from the analysis (61 pupils at 7 schools). Summary statistics of the remaining 535,829 pupils in 3,017 schools for the three variables detailed above are given in Table 2, below:

Table 2 – Summary statistics for 2013 NPD variables

		Mean	Std. Dev	Min	Max
Pupil-level (n = 535,829)	Best 8 Score Plus Bonus	428.3	88.8	0	580
	Best 8 Value-Added Score	-0.1	67.1	-537.7	306.1
	Key Stage 2 Average Point Score	27.9	3.9	3	35.7
School-level† (n = 3,017)	Best 8 Score Plus Bonus	428.1	35.4	302.1	564.1
	Best 8 Value-Added Score	0.2	22.0	-90.7	68.2
	Key Stage 2 Average Point Score	27.8	1.5	23.4	33.5

†School mean of each score

4.4 Simulation Design

The following simulation is designed to estimate the extent to which random errors in KS2 scores translate into errors in the school-level value-added scores. Value-added scores are comprised of school effects and what Amrein-Beardsley (2014, 38) call 'construct irrelevant variance' (CIV), a term which denotes all sources of variation that are not attributable to schools including all biases and errors. As discussed above, it is difficult to know to what extent any given scores are comprised of CIV. When one adds error to actual value-added scores as part of a simulation, this is added to any error already present. Conversely, simulations that adjust for error will under- or over-correct depending on the approach and formulation, again leading to estimates containing a mixture of added and pre-existing (and generally unobserved) CIV. These approaches are both potentially informative. The alternative taken here, however, is to observe the effect of error independently of all other sources of variation while still using realistic data. The steps taken to achieve this are as follows:

The first step was to create a deterministic but realistic KS2-KS4 attainment relationship. To do this, the actual KS2 data and school memberships from the 2013 pupil-level KS4 NPD data extract were used; the KS2 scores for these pupils were recorded in 2008 (see Table 2 above for summary statistics). Then, a slightly simplified version of the model used for the official 2013 DfE VA measure was estimated (see equation 1 below) to produce an expected KS4 score for each KS2 score. This expected score notionally represents the KS4 distribution in the absence of any value-added, biases or error (value-added is calculated as the actual score minus this expectation). So, for the purposes of this simulation, it is imagined that these expected scores *are* the true KS4 scores of the pupils in question and schools are not differentially effective. This disregards all residual variation comprised of value-added and any CIV. From this point, error was added to the KS2 scores and the extent to which this affected

the estimated value-added scores was calculated. As school value-added should be zero by design, if there is any value-added at all, it can be directly attributed to measurement error, thereby giving information about the scale and nature of the impact of various levels of random measurement error in this context. Put simply, this design poses the hypothetical question: 'In a world where there is no school effect or any other non-school influence on performance, how much spurious school value-added can arise just from measurement error in the baseline attainment measure?'

In sum, this simulation has been designed to demonstrate the effects of adding error to an initial error-free scenario while keeping as close as possible to the actual English data to obtain realistic estimates of the effects. The design has the following characteristics: 1) the KS2 scores, pupil numbers and school memberships are the same as the actual data 2) the relationship between KS2 and KS4 performance is deterministic (i.e. there is no value-added or other CIV), and c) the constructed value-added-free KS4 distribution is highly similar to the actual KS4 distribution. The mean values of the expected and actual distributions were identical to 4 decimal places and, as would be expected with value-added and CIV removed, the standard deviation of the former was smaller at 58 compared to 89 (cf. Table 2).

4.5 Model Specification

The following model is based on the replica of the official 2013 VA measure (DfE 2013, 18). To simplify the analysis, the maths deviation and English deviation scores were dropped from the value-added model. The effect of omitting these is minimal: the correlation between the expected scores with and without is 0.996 (3dp). The advantage of dropping these is that, without these deviations, the model produces a one-to-one correspondence between KS2 and KS4 scores. The following model was used to create KS4 expectations to be used as simulated KS4 scores with value-added removed. This model is also used for subsequent estimates of any spurious value-added created by introducing error:

1) $Best8ScorePlusBonus_{ij} = \beta_0 + \beta_1 KS2APS_{ij} + \beta_2 KS2APS^2_{ij} + \beta_3 KS2APS^3_{ij} + \varepsilon_{ij}$

Where the subscript j denotes schools;

The subscript i denotes pupils;

$Best8ScorePlusBonus_{ij}$ is the 2013 official KS4 attainment measure;

β_0 is a constant intercept term;

ε_{ij} is the model residual, which includes any value-added and CIV;

$KS2APS_{ij}$ is the Key Stage 2 average point score.

This was estimated within a multilevel framework in which random effects at pupil (e_{ij}) and school level (u_j) could be obtained from the model residuals, as follows:

2) $\varepsilon_{ij} = u_j + e_{ij}$

The analysis was also conducted using the value-added method from the new Progress 8 measure (DfE 2016), following the details given in Burgess and Thomson (2013b). As Burgess and Thomson (2013b) find, the simpler Progress 8 method produces estimates that are almost identical to value-added models such as the one specified above. In the present study, the expected scores produced using the Progress 8 method and those using the model described above had a pupil-level correlation of 0.999 (3dp). The advantage of reporting the latter is that this more closely simulates the official value-added model from 2013 and the use of regression-based models allows the reporting of model outputs for easier comparison with other studies. All results are almost identical with both approaches and are so are directly relevant to the current Progress 8 measure.

4.6 Simulated Error Rates

A key methodological choice in this simulation is the size of the error introduced. The seriousness of attenuation bias and error will be highly dependent on measurement reliability, which can vary considerably. The Performance Indicators in Primary Schools (PIPS) data used in Harker and Tymms (2004) is reported to have a test re-test reliability of 0.96 and, as they note, is 'constructed specifically as a predictor of later academic achievement' (see 191). Yet, many measures will fall far short of this. In a English policy context, reliability studies conducted by Ofqual, the English exam regulator (Opposs and He 2011, 10–12), estimated KS2 Reading parallel forms reliabilities of 0.85 and 0.81 when comparing to anchor tests and 'live' tests taken a few weeks after the first test and between 0.81 and 0.88 for KS2 Science when comparing pre-tests and live tests over several years. The 'large', 'medium' and 'small'

errors in the present study (below) are designed to be approximately in line with this range (0.81 to 0.96). Of course, the designation of these errors rates as small, medium and large is uncertain and contestable, hence the value in presenting results at three levels of error. Also, to reiterate a point made earlier, it is likely that errors are non-random and differ across the score scale. This is another possible source of bias not considered here.

In the present simulation, three levels of error were added to the KS2 scores. To do this, 3 normally-distributed random variables were created. These error distributions all had a mean of 0. The standard deviations of these error variables and their correlation with the original scores are given in Table 3, below:

Table 3 – Random error introduced into KS2 scores during simulation

	Error Standard Deviation on National Curriculum Point Scale	Correlation between true and error-adjusted scores
Small error	1 NC point (1/6 of a NC level).	0.97
Medium error	2 NC points (1/3 of a NC level).	0.89
Large error	3 NC points (1/2 of a NC level).	0.79

These errors were added to the actual KS2 scores to create three new KS2 score variables, one for each level of error. As described above, for the first application of the model, the model expectations were saved and subsequently treated as the actual KS4 scores, creating a deterministic KS2-KS4 relationship with all value-added removed. After introducing the KS2 errors, the same model was used to estimate school value-added (u_j), this time using the KS2 scores with added measurement error as baseline prior attainment measures. If the pupil-level errors cancel out at school-level and there is no systematic bias, the school-level value-added estimates from these latter models should be approximately zero, indicating that the measurement error has not affected the KS2-KS4 relationship or school scores.

One final noteworthy point relates to the decision to introduce error at KS2 but not KS4. Measurement error in the baseline score has two effects: First, the direct effect of the error leads to errors in pupils' value-added scores. This effect is straight-forward and affects the pupils in question according to the extent of the error. When these are aggregated by school or national cohort, given the assumption that they are random, the impact will be considerably smaller. The second effect comes from the attenuation bias which shrinks the expected scores towards the mean. Even pupils with no individual error whatsoever will have their value-added increased in either a positive or negative direction by attenuation bias. For the final KS4 scores

only the first, less serious, problem applies. This was tested empirically in a separate analysis (not reported) but also applies in theory (see Greene 2000, 84 who also presents a theoretical treatment of attenuation bias).

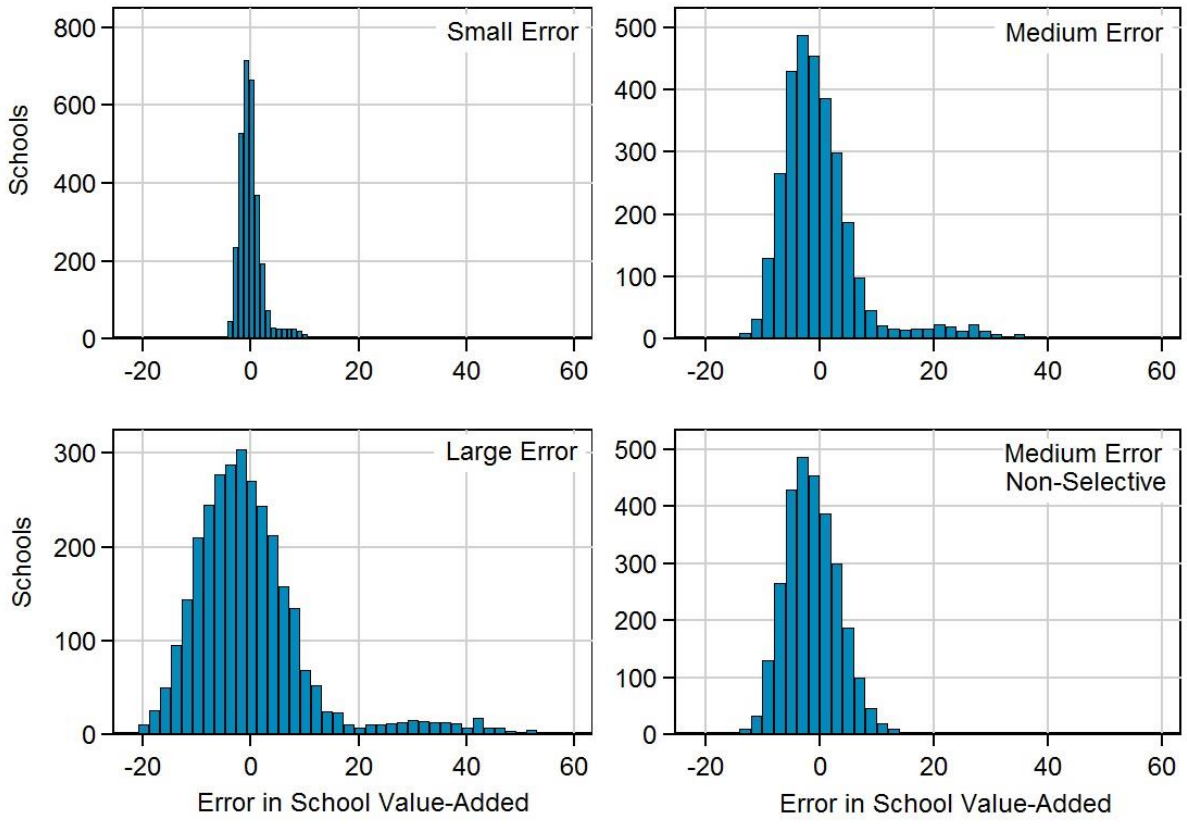
5. Results

5.1 Simulation Results

The value-added model specified in Section 4.5, above, was calculated to estimate the value-added using the true KS2 scores and then again for each level of error. The scale for the school value-added corresponds with the Best 8 Score measure, where a score of 6 equates to 1 additional GCSE grade per pupil at the school across the 'Best 8' GCSE exams with Maths and English double-weighted (see Table 2). Model outputs are given in Table A1, Appendix A.

The amount of error translating into school-level value-added from each error level is shown in a series of graphs in Figure 2 below. The first three graphs correspond to the small, medium and large error rates. The final graph is for the medium error rate with academically selective schools removed (see below for further details).

Figure 2 – Error introduced into KS2-4 school-level value-added by KS2 error rate



These graphs illustrate two important findings: First, school-level error rates are substantial, in particular for the medium and large levels of pupil-level error. Second, the first three plots have a clear right-hand tail suggesting a systematic bias. The final histogram shows the rates of school-level error when excluding schools with mean intake KS2 scores of 31 and above (level 5C at KS2) from the plot (but not the original analysis). This removes the 161 most highly selective schools in relation to intake prior attainment, overwhelmingly comprised of academically selective 'grammar' schools. This final histogram suggests that the systematic bias in the other graphs is linked with school attainment composition; in other words, a phantom compositional effect on attainment (see earlier). These two findings are considered in turn.

5.2 Finding 1 - Substantial Error Rates

These results have used three error levels to show the magnitude of error which can translate from pupil-level error to school-level scores at each. Summary statistics at each level of error, for all schools and when excluding selective schools are given in Table 4, below:

Table 4 – Summary statistics of the impact of the three KS2 error rates on KS2-4 school value-added with and without the 161 most selective schools removed

All Schools

	Schools	Mean	Std. Dev.	Min	Max
Small Error	3017	0.0	2.4	-5.3	15.0
Medium Error	3017	0.0	7.1	-13.9	36.3
Large Error	3017	0.0	11.2	-20.8	56.7

Selective Schools Removed

	Schools	Mean	Std. Dev.	Min	Max
Small Error	2856	-0.4	1.6	-5.3	5.4
Medium Error	2856	-1.3	4.6	-14.0	16.1
Large Error	2856	-2.0	7.4	-20.8	28.5

These errors can be put in context of the overall Best 8 value-added school-level distribution which spans from around -90.7 to 68.2 (see Table 2, above). These results suggest that substantial error rates can translate from pupil-level scores to school-level scores. Even when disregarding the right-hand tail, the school-level error distributions ranged from approximately 5, 15 and 25 points above and below zero. This corresponds to an average of approximately 1, 2-3 and 4 GCSE grades *per pupil* across the best 8 GCSE subjects.

5.3 Finding 2 - Systematic Bias

The simulation was repeated using another set of random errors, specified as before. These analyses all gave the same substantive result. The correlation between school-level errors in small, medium and large error conditions for the original and repeated analyses were 0.83, 0.94, 0.97 respectively. This means that the same schools can expect to be systematically (dis)advantaged under these conditions based on the characteristics of their intake rather than this being largely driven by chance.

As demonstrated in the final graph in Figure 2, above, the right-hand tail is largely eliminated by excluding the 161 most selective schools according to intake mean KS2 scores. This is a very interesting result given the apparent 'grammar school effect' that is often found using English data (Sutton Trust et al. 2008). Note that there are currently 163 schools listed as grammar schools in England (Bolton 2015). In this simulation, the KS4 scores have been replaced with a *deterministic* score, with the value-added removed. Yet, despite this, with the introduction of random error, selective schools still emerge as having a disproportionate number of pupils with a positive value-added which, in this case, is known to be entirely spurious. With the introduction of a small, medium and large amount of pupil-level error, the mean errors at school-level for just these 161 schools are 7.4, 22.8 and 35.1 respectively. The 'phantom grammar school effect' of 22.8 for the medium level of error is especially remarkable given that a recent estimate of the grammar school effect in official English value-added data (2015) of approximately 25 points (Morris and Perry, 2016).

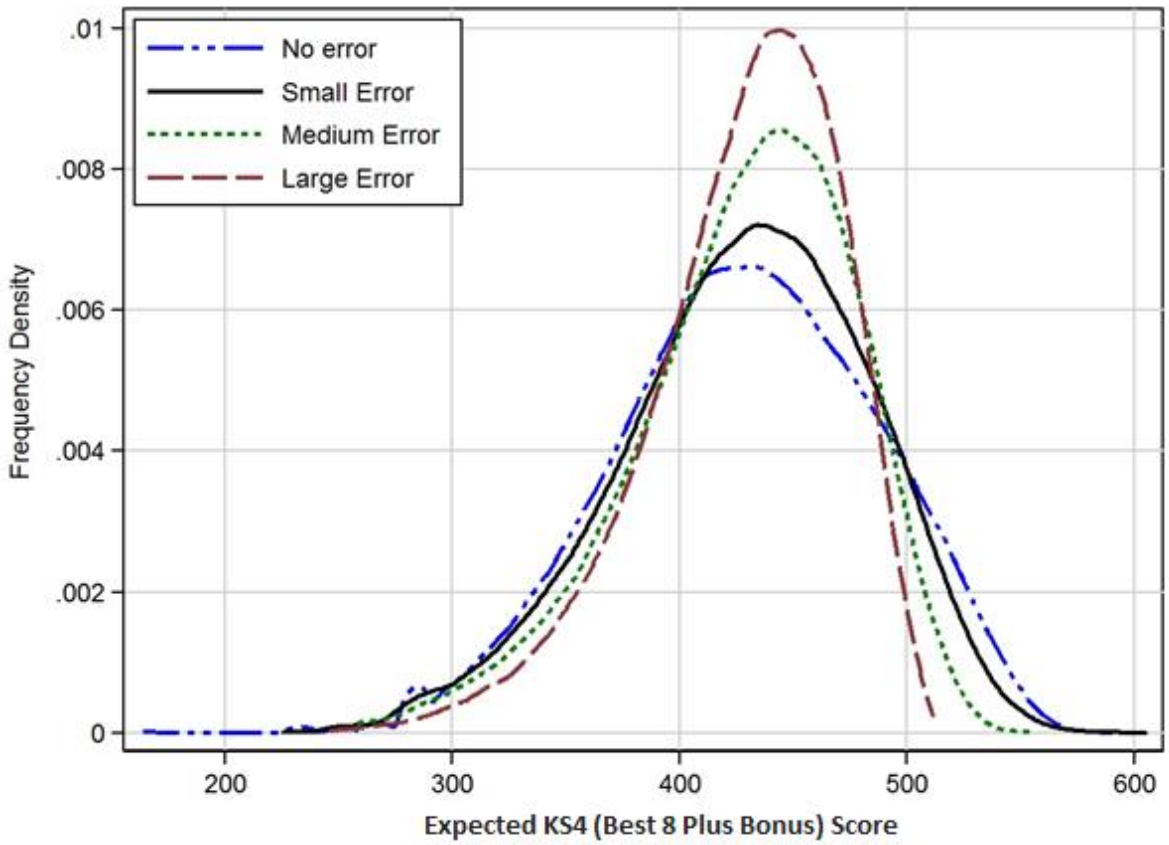
Examination of the model outputs and expectations clearly demonstrated that the systematic bias was caused by attenuation bias. This is less apparent from the model output (as given in Appendix A) because the specification included squared and cubed terms and this can make the attenuation less clear due to the weight being shifted between the original, squared and cubed terms. The attenuation bias is very clearly apparent, however, from summary statistics of the model expectations at each level of error, given in Table 5, below:

Table 5 – Summary statistics of pupil-level KS4 expectations

	Mean	Std. Dev.	Min	Max
Expectation with small error	428.3	54.9	225.6	604.8
Expectation with medium error	428.2	48.8	258.3	554.8
Expectation with large error	428.2	42.6	246.1	511.7

Table 5 shows that, while the mean expectation remained the same, as more error was introduced, the standard deviation of the expectations fell. This shrinkage of the expectations towards the mean value can also be seen in Figure 3, below, which gives the frequency distribution for attainment expectations at each level of error.

Figure 3 – Expected KS4 values with increased amounts of error



To see the effect of the error on the pupil- and school-level scores across the pupil KS2 and school mean KS2 attainment ranges, Figure 4, below, gives the spurious value-added scores at both pupil- and school-level against both the actual KS2 scores (left-hand column) and the KS2 scores with error (right). In a practical setting, only the latter (right) are observed and there would be an apparent compositional effect (bottom right) despite no observed relationship at pupil-level (top right).

Figure 4 – Unobserved and observed medium error at pupil- and school-level

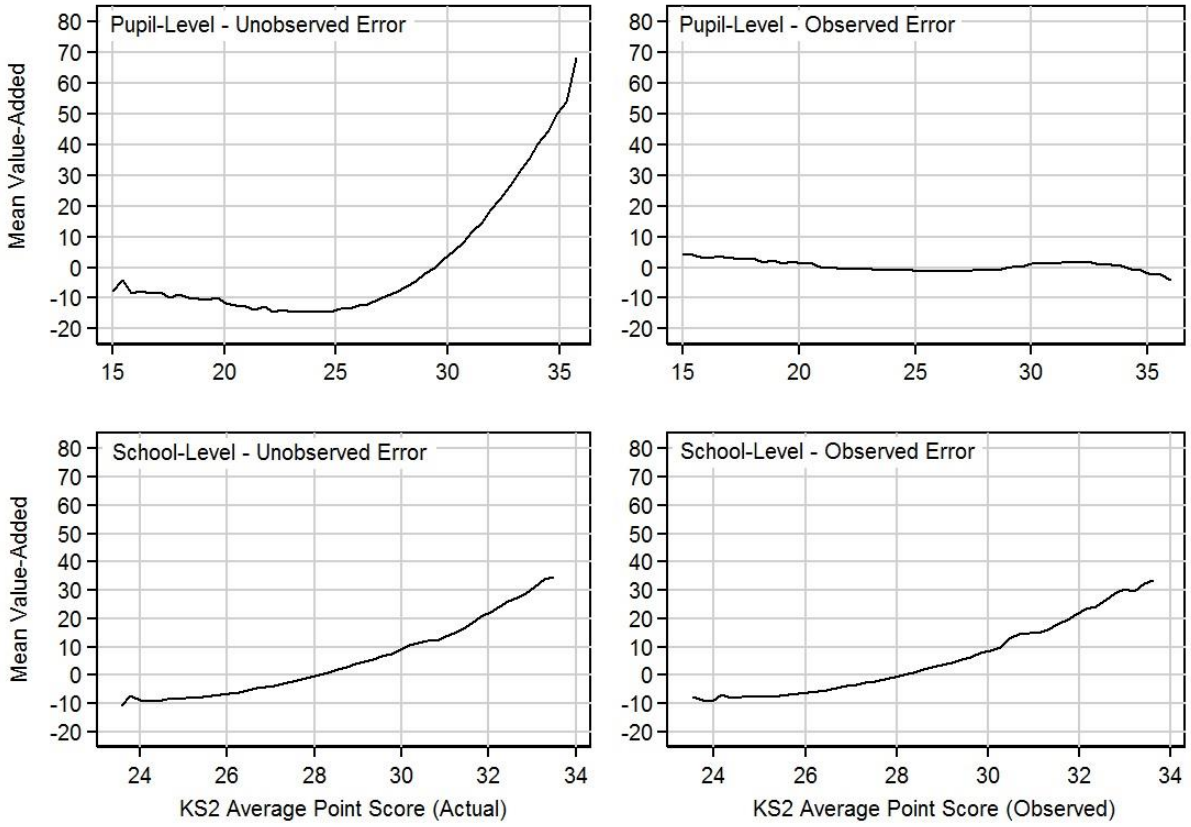


Figure 4 shows the consequences of the attenuation of the expectations illustrated earlier in Figure 3. When a pupil or school has a KS2 score of about the mean of 28 (cf. Table 2), the expected value of the value-added score is zero. For pupils or schools above or below the mean, there is a positive or negative bias approximately in proportion to their distance from the mean, tailing off at the lower end of the pupil-level distribution. It is also noteworthy that the observed and unobserved school-level biases are almost identical, suggesting that the pupil-level error that obscures the relationship at pupil-level largely cancels out, revealing a close approximation of the true relationship at school-level (Harker and Tymms 2004, 195).

There are several methodologically important points to be taken from this: first, that correcting for pupil attainment at pupil-level is not necessarily sufficient to correct for it at school-level. This has implications for the English Progress 8 approach which does not include school-level variables. Second, in line with previous studies (e.g. Harker and Tymms 2004; Televantou et al. 2015), this study finds that phantom compositional effects can arise from pupil-level measurement error. The focus of previous studies, however, has been on the methodological difficulties of estimating compositional effects; what has not been made clear but is apparent in Figure 4 is that the value-added score can have a systematic, unobserved bias at pupil-level. Clearly this has implications when value-added scores are used at pupil-level

such as target setting and at school-level because a school's value-added score will be affected by the profile of its individual composition as well as its average. Improved accuracy in pupil-level predictions could be achieved by adjusting pupils' expectations in the direction of the cohort mean expectation.

5.4 Separating Direct and Indirect (Attenuation) Effects of Error

As this study has the 'true' KS2 scores, it is possible to observe and remove the systematic biases shown in Figure 4 to examine the extent to which the school-level errors shown in Figure 2 are systematic and to what extent they are the result of random pupil-level errors failing to cancel out at school-level. This is one advantage of the methodological approach taken: with value-added and CIV removed, any observed value-added is known to be spurious. Details of this correction, using a multiple regression analysis, are given in Appendix B.

Table 6, below, shows summary statistics of the remaining spurious school value-added after removing systematic biases shown in Figure 4. This shows small rates of school-level error which will be negligible for most schools.

Table 6 – Error variance with all systematic error variance corrected for

	Mean	Std. Dev.	Min	Max
Small Error	0.0	0.5	-2.1	3.1
Medium Error	0.0	1.0	-3.9	4.4
Large Error	0.0	1.2	-5.8	4.5

This result is valuable theoretically to estimate the extent to which error was systematic. For practical purposes, however, the true scores are not available, only the observed (error-adjusted) scores. One final correction approach tested, therefore, was to regress the school-level (spurious) value-added scores on the school-level error-adjusted KS2 scores. A squared term was also included due to slight non-linearity (see Figure 4) and their standard deviation, as was done in the former English contextualised value-added measure (2005-2010) (Evans 2008). The KS2 scores at small, medium and large error rates explained (R^2) 80%, 89%, 90% of the school-level error variance. This suggests that a school-level control for prior attainment will largely, but not entirely, correct the problem of attenuation bias in this context for purposes of school-level inferences. Despite this vast improvement, however, appreciable amounts of error remain in the scores (see Table 7). With this school-level correction, error rates are considerably smaller but still of some concern at the extremes, reaching 1, 2 or even 3 GCSE grades per pupil depending on the level of error. With average prior attainment corrected, these errors reflect differences in schools' prior attainment distributions.

Table 7 – Error variance with observed systematic error variance corrected for

	Mean	Std. Dev.	Min	Max
Small Error	0.0	1.1	-4.3	4.6
Medium Error	0.0	2.4	-9.2	9.4
Large Error	0.0	3.5	-15.7	18.9

6. Theoretical and General Implications

In methodological debates reviewed in the opening section, Reynolds et al. (2012, 116) held that errors 'tend to be randomly distributed' and so have little impact on the school-level scores, whereas Gorard (2011a, 18) argued that this is unlikely and that 'it is unfair and unethical' to assume this is the case. The remarkable result from this study is that *even if* pupil-level errors are random (across pupils, time and schools), these can still translate into serious school-level errors if attenuation bias is not addressed.

There are several important considerations pertaining to whether this problem will be apparent more generally. First, it has been discussed that baseline measures will vary considerably in their reliability and the simulation exhibits marked differences in attenuation bias at different error rates. It is possible to increase measurement reliability to some degree by using advanced techniques such as those examined in Televantou (2015) and Pokropek (2014). Such correction approaches take advantage of the ability to break down tests into 'parcels of items' as multiple indicators thereby improving their predictive value and correcting for some portion of the measurement error (Televantou et al. 2015, 76).

The second consideration is the extent to which corrections are made for school-level prior attainment. Note that this is to assume that school-level relationships reflect error rather than a genuine compositional effect, an assumption which the literature on compositional effects suggests is reasonable (see Section 2.3). Many educational effectiveness studies have included school-level controls and will therefore be less affected by the problem. It has been found here that, for smaller error rates with school-level relationships controlled (see Section 5.4), error translating into school-level scores was modest.

Third is the issue of outputs and use of value-added models. An important point to note is that attenuation bias is not a peculiarity of certain regression equations or value-added models and will also affect the English 'Progress' scores. This has been found empirically in this study using simulated data and also holds in theory (also see studies finding attenuation bias from other contexts, Section 2.4 above). Nonetheless, differences in the models and estimation techniques as well as the use of the estimates will influence the severity of the

problem. For example, the English Progress 8 measure is used to make judgements about individual schools, where biases will have serious consequences: each individual score matters. In contrast, researchers tend to look at groups of schools or the entire sample for more low-stakes purposes. In the context of an educational effectiveness study it is generally the effectiveness factors associated with value-added that are the main focus rather than individual schools' scores. Problems may arise wherever factors are associated with average ability, particularly when there are no school-level controls. Any practice that is more or less common in schools according to average ability will be identified as an effectiveness factor.

Before looking at the policy implications of this study, it is valuable to comment on how this study's findings relate to the studies of measurement error reviewed in Section 2. These concluded that the effects of error are 'confined to the fixed part of the model' (Goldstein, Kounali, and Robinson 2008, 256), where the implication is that error does not affect the residual, or 'random', school effects. There is something potentially very misleading about this conclusion, certainly when it comes to the drawing implications of measurement error in a policy context. A consistent finding from Woodhouse et al. (1996), Goldstein, Kounali, and Robinson (2008) and Ferrão and Goldstein (2009) is that adjusting for measurement error increases the coefficient on prior attainment. In other words, it offsets attenuation bias. This puts the results but not the conclusions of these studies in line with the present study. It is a curious conclusion however that there can be substantial changes to the coefficient values of vital control variables such as prior attainment *without* affecting the random component of the model. As prior attainment coefficients change, so must school scores according to the prior attainment of their intake. This highlights the differences in application of value-added (see above) as well as the limitation of defining 'school effect' in terms of relative variance and the use of intraclass correlation coefficients to measure it. Moreover, the use of correlations between original and corrected scores (as in Ferrão and Goldstein 2009) is also potentially misleading: a partial adjustment of a systematically biased score will still be highly correlated if a substantial proportion of the school-level value-added is due to systematic error. This highlights the advantage of the present simulation approach which can unambiguously attribute changes in school scores to error and compare the results with a constructed error-free estimate.

7. The Effects and Seriousness of Error in the English Accountability Context

7.1 Key Policy-Relevant Findings

This study has reached several conclusions about the likely effect of random measurement error on the English school value-added scores:

First, without correction for attenuation bias, levels of error in KS2 prior attainment score in line with reasonable expectations and existing reliability evidence translate to serious errors in school-level value-added scores. For selective grammar schools at a medium error rate, this bias can be as high as 36 'Best 8' KS4 points, or 6 GCSE grades per pupil over the best 8 GCSE scores. Error rates for non-selective schools scores were as high as 2 to 3 GCSE grades per pupil.

Second, these errors were found to primarily stem from attenuation bias in the pupil-level value-added. This bias was only observable when scores were aggregated to school-level, linking to a strand of research on 'phantom' compositional effects. The estimates for the English data suggest that the school-level bias is largely observable. This means that it is possible to correct the bias to a large degree in this context using simple school-level controls.

Third, analysis of compositional effects and grammar school effects in 2004-2016 English data suggest that changes to recent KS4 measures, earlier KS2 measures or both have inflated the apparent compositional effect. Even the largest estimates of grammar school effects in more recent years are in line with the 'phantom grammar school effects' found at the medium error rate in the simulation analysis presented here. This finding and the analysis of the 2004-2016 English data very strongly suggest that the compositional effects and grammar school effects found in the English accountability data are largely or wholly spurious.

7.2 The Effect of Policy Decisions

The policy decision in 2010 to disregard contextual variables (including school-level controls), as well as raising concerns over biases relating to pupil background (Perry, 2016), has left the English value-added measures wide open to attenuation bias. The errors found in the present study, in terms of their size and their systematic tendency, suggests that this decision was mistaken. These results have increased importance due to a debate taking place about the merits of grammar schools (Morris and Perry 2016), a debate predicated on their superior effectiveness.

Limiting the Progress 8 measure to pupil-level prior attainment controls was a specific requirement given for the measure by the DfE. The resultant potential for bias, (albeit not attenuation bias specifically) was clearly noted in the initial report examining the options

(Burgess and Thomson 2013a). There is certainly merit in favouring simpler models, especially when estimates are highly similar; trying to eliminate *all* threats to validity tends to require extensive data, more complex models and highly-technical adjustments. This lacks transparency and brings considerable practical drawbacks for what are often small benefits to validity (Kelly and Downey 2010). With the new 'Progress 8' measure, however, English policy makers are at the other extreme, having stripped back the measure to the level of basic arithmetic, disregarded all contextual and school-level variables and thereby precluding the chance to deal with school-level biases such as found here.

The continued use of school value-added performance measures that do not address the problem of attenuation bias is untenable, especially when there are clear steps that can be taken to at least partly mitigate its effects. There are several possible options such as including school-level variables in a value-added model; adjusting scores at school-level in a second step; using various measurement error correction techniques to minimise measurement error (see Televantou et al. 2015); and designing KS2 examinations (or separate baseline tests) specifically to maximise their reliability and their ability to predict the KS4 scores.

7.2 The Wider Problem

An issue that runs through this entire study is the assumed randomness of measurement error. There is no reason to assume that measurement errors are wholly or even largely random (Gorard 2011a). The simulation presented here considers only the effect of random error and considers only one aspect of the problem. Therefore, the impact of error is potentially considerably worse than presented here. Furthermore, it is also worth returning to the point made in the introduction to this paper that, in addition to measurement error, a major threat to the validity of school value-added scores is omitted variable bias. The most recent estimates suggest that over a third of the variance in the KS2-4 English value-added measure is explained by a small handful of pupil background factors, notably disadvantage, English language status and, as highlighted in the present study, prior attainment composition (Perry 2016).

It is also important to consider the impact of error and bias on the consistency and stability of school value-added scores (Perry 2016). Systematic biases such as phantom compositional effects or those stemming from omitted variable bias will have a *stabilising* effect on value-added scores, with the same schools being (dis)advantaged over time. If these are removed, the resultant scores are likely to become increasingly unstable (Dumay, Coe, and Anumendem 2014). With all such biases stripped away from the measure, how stable will the

resulting scores be? The evidence pertaining to the former CVA measure suggests that stability may be unacceptably low (Leckie and Goldstein 2011; Gorard, Hordosy, and Siddiqui 2012).

Such measurement unreliability in the context of accountability has serious consequences. By international accountability standards, England is characterised as having a particularly 'high-pressure system' (Altrichter and Kemethofer 2015, 50). The results of high-stakes inspections, where schools are closed or expanded and school leaders are rewarded or fired, are heavily influenced by such performance measures. With the new Progress 8 measure being used as a headline measure of school performance for the first time, the consequences of error and bias have never been higher.

Acknowledgements

The author would like to thank Peter Davies and Stephen Gorard for their support and advice over the course of the author's doctoral research programme from which this paper arises. Thanks also go to the ESRC for funding the research (ref: 1185067).

References

AERA. 2015. "AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs." Online: AERA. http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html.

Altrichter, Herbert, and David Kemethofer. 2015. "Does Accountability Pressure through School Inspections Promote School Improvement?" *School Effectiveness and School Improvement* 26 (1). Routledge: 32–56. doi:10.1080/09243453.2014.927369.

Amrein-Beardsley, Audrey. 2014. *Rethinking Value-Added Models in Education: Critical Perspectives on Tests and Assessment-Based Accountability*. Routledge.

Bolton. 2015. "Grammar School Statistics." House of Commons Library. doi:10.1093/nq/s1-IV.105.345-b.

Boonen, Tinneke, Sara Speybroeck, Jerissa Bilde, Carl Lamote, Jan Damme, Patrick Onghena, Jan Van Damme, and Patrick Onghena. 2014. "Does It Matter Who Your Schoolmates Are? An Investigation of the Association between School Composition, School Processes and Mathematics Achievement in the Early Years of Primary Education." *British Educational Research Journal* 40 (3): 441–66.

Burgess, Simon, and Dave Thomson. 2013a. "Key Stage 4 Accountability: Progress Measure and Intervention Trigger." http://www.bristol.ac.uk/cubec/portal/: BUBeC, University of Bristol.

———. 2013b. "Key Stage 4 Accountability: Progress Measure and Intervention Trigger, Technical Annex: Techniques for Producing an Unbiased National Pupil Progress Line." http://www.bristol.ac.uk/cubec/portal/: BUBeC, University of Bristol.

DfE. 2013. "A Guide to Value Added Key Stage 2 to 4 in 2013 School Performance Tables & RAISEonline." Edited by Department for Education. DfE.

http://www.education.gov.uk/schools/performance/2013/secondary_13/KS2-4_Performance_Tables_General_VA_Guide_2013_FINAL.pdf.

———. 2016. “Progress 8 Measure in 2016, 2017, and 2018.” Department for Education. https://www.gov.uk/government/uploads/system/uploads/attachment%7B_%7Ddata/file/497937/Progress-8-school-performance-measure.pdf.

———. 2017. “Progress 8 and Attainment 8.” https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/583857/Progress_8_school_performance_measure_Jan_17.pdf.

Dieterle, Steven, Cassandra M Guarino, Mark D Reckase, and Jeffrey M Wooldridge. 2015. “How Do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value Added.” *Journal of Policy Analysis and Management* 34 (1): 32–58.

Dumay, Xavier, Rob Coe, and Dickson Nkafu Anumendem. 2014. “Stability over Time of Different Methods of Estimating School Performance.” *School Effectiveness and School Improvement* 25 (1). Routledge: 64–82. doi:10.1080/09243453.2012.759599.

Evans, Helen. 2008. “Value-Added in English Schools.” London: Dept for Children, Schools, Families. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.9363&rep=rep1&type=pdf.

Ewijk, Reyn van, and Peter Sleegers. 2010. “Peer Ethnicity and Achievement: A Meta-Analysis into the Compositional Effect.” *School Effectiveness and School Improvement* 21 (3). T&F Informa UK Ltd: 237–65. http://www.swetswise.com/swoc-web/link/access_db?issn=1744-5124&vol=00021&iss=00003&page=237&year=2010.

Ferrão, Maria Eugénia, and Harvey Goldstein. 2009. “Adjusting for Measurement Error in the Value Added Model: Evidence from Portugal.” *Quality & Quantity* 43 (6): 951–63.

Frost, Chris, and Simon G Thompson. 2000. “Correcting for Regression Dilution Bias: Comparison of Methods for a Single Predictor Variable.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163 (2): 173–89. doi:10.1111/1467-985X.00164.

Gibbons, Stephen, and Shqiponja Telhaj. 2012. “Peer Effects: Evidence from Secondary School Transition in England.” 6455. Discussion Paper Series, Forschungsinstitut Zur Zukunft Der Arbeit. http://ftp.iza.org/dp6455.pdf.

Goldstein, Harvey, Daphne Kounali, and Anthony Robinson. 2008. “Modelling Measurement Errors and Category Misclassifications in Multilevel Models.” *Statistical Modelling* 8 (3): 243–61.

Gorard, Stephen. 2006a. “Is There a School Mix Effect?” *Educational Review* 58 (1): 87–94.

———. 2006b. “Value-Added Is of Little Value.” *J. Educ. Policy* 21 (2): 235–43.

———. 2010. “Serious Doubts about School Effectiveness.” *British Educational Research Journal* 36 (5): 745–66.

———. 2011a. “Comments on ‘The Value of Educational Effectiveness Research.’” *BERA Conference*. BERA. http://beraconference.co.uk/index.php/download_file/view/42/75/.

———. 2011b. “Doubts about School Effectiveness exacerbated–By Attempted Justification.” *Research Intelligence* 114 (26): Q8.

Gorard, Stephen, R Hordosy, and N Siddiqui. 2012. “How Unstable Are ‘School Effects’ Assessed by a Value-Added Technique?” *International Education Studies* 6 (1): p1.

Greene, William H. 2000. *Econometric Analysis (International Edition)*. Pearson Education Ltd. New Jersey USA: Prentice Hall.

Harker, Richard, and Peter Tymms. 2004. “The Effects of Student Composition on School Outcomes.” *School Effectiveness and School Improvement* 15 (2): 177–99. http://www.tandfonline.com.ezproxyd.bham.ac.uk/doi/pdf/10.1076/sesi.15.2.177.30432.

He, Qingping, Malcolm Hayes, and Dylan Wiliam. 2013. “Classification Accuracy in Key Stage 2 National Curriculum Tests in England.” *Research Papers in Education* 28 (1). Routledge: 22–42. doi:10.1080/02671522.2012.754225.

Hutchison, Dougal. 2007. “When Is a Compositional Effect Not a Compositional Effect?” *Quality & Quantity* 41 (2): 219–32. doi:10.1007/s11135-007-9094-2.

Kane, Michael T. 2013. “Validating the Interpretations and Uses of Test Scores.” *Journal of Educational Measurement* 50 (1): 1–73. doi:10.1111/jedm.12000.

Kelly, Anthony, and Christopher Downey. 2010. “Value-Added Measures for Schools in England: Looking inside the ‘black Box’ of Complex Metrics.” *Educational Assessment, Evaluation and Accountability* 22 (3): 181–98. doi:10.1007/s11092-010-9100-4.

Lavy, Victor, Olmo Silva, and Felix Weinhardt. 2012. “The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools.” *Journal of Labor Economics* 30 (2): 367–414.

Leckie, George, and Harvey Goldstein. 2011. “Understanding Uncertainty in School League Tables.” *Fiscal Studies* 32 (2). Blackwell Publishing Ltd: 207–24. doi:10.1111/j.1475-5890.2011.00133.x.

Liu, Hongqiang, Jan Van Damme, Sarah Gielen, and Wim Van Den Noortgate. 2015. “School Processes Mediate School Compositional Effects: Model Specification and Estimation.” *British Educational Research Journal* 41 (3): 423–47. doi:10.1002/berj.3147.

Marks, Gary N. 2015. “Are School-SES Effects Statistical Artefacts? Evidence from Longitudinal Population Data.” *Oxford Review of Education* 41 (1). Routledge: 122–44. doi:10.1080/03054985.2015.1006613.

McCaffrey, Daniel F, Katherine E Castellano, and J R Lockwood. 2015. “The Impact of Measurement Error on the Accuracy of Individual and Aggregate SGP.” *Educational Measurement: Issues and Practice* 34 (1): 15–21. doi:10.1111/emip.12062.

Morris, Rebecca, and Thomas Perry. 2016. “Reframing the English Grammar Schools Debate.” *Educational Review* 69 (1). Routledge: 1–24. doi:10.1080/00131911.2016.1184132.

Muijs, Daniel, Tony Kelly, Pam Sammons, David Reynolds, and Chris Chapman. 2011. “The Value of Educational Effectiveness Research: A Response to Recent Criticism.” *Research Intelligence* 114: 24–25.

Nash, Roy. 2003. “Is the School Composition Effect Real?: A Discussion With Evidence From the UK PISA Data.” *School Effectiveness and School Improvement* 14 (4): 441–57. doi:10.1076/sesi.14.4.441.17153.

Newton, Paul E. 2013. “Ofqual’s Reliability Programme: A Case Study Exploring the Potential to Improve Public Understanding and Confidence.” *Oxford Review of*

Education 39 (1). Routledge: 1–21. doi:10.1080/03054985.2012.760285.

Opposs, Dennis, and Qingping He. 2011. "The Reliability Programme: Final Report." http://dera.ioe.ac.uk/2568/1/11-03-16-Ofqual-The-Final-Report.pdf.

Perry, Thomas. 2016. "English Value-Added Measures: Examining the Limitations of School Performance Measurement." *British Educational Research Journal* 42 (6): 1056–80. doi:10.1002/berj.3247.

Pokropek, Artur. 2014. "Phantom Effects in Multilevel Compositional Analysis: Problems and Solutions." *Sociological Methods & Research*, 1–29. doi:10.1177/0049124114553801.

Reynolds, David, Christopher Chapman, Anthony Kelly, Daniel Muijs, and Pam Sammons. 2012. "Educational Effectiveness: The Development of the Discipline, the Critiques, the Defence, and the Present Debate." *Effective Education* 3 (2). Routledge: 109–27. doi:10.1080/19415532.2011.686168.

Reynolds, David, Pam Sammons, Bieke De Fraine, Jan Van Damme, Tony Townsend, Charles Teddlie, and Sam Stringfield. 2014. "Educational Effectiveness Research (EER): A State-of-the-Art Review." *School Effectiveness and School Improvement* 25 (2). Routledge: 197–230. doi:10.1080/09243453.2014.885450.

Snijders, Tom A B, and Roel J Bosker. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London: Sage.

Sutton Trust, Sutton Trust, Sutton Trust, Sutton Trust, Robert Coe, Karen Jones, Jeff Searle, et al. 2008. "Evidence on the Effects of Selective Educational Systems." *A Report for the Sutton Trust. Durham: CEM Centre, University of Durham, for the Sutton Trust*. CEM Centre, Durham University, UK www.cemcentre.org.

Televantou, Ioulia, Herbert W Marsh, Leonidas Kyriakides, Benjamin Nagengast, John Fletcher, and Lars-Erik Malmberg. 2015. "Phantom Effects in School Composition Research: Consequences of Failure to Control Biases due to Measurement Error in Traditional Multilevel Models." *School Effectiveness and School Improvement* 26 (1). Routledge: 75–101. doi:10.1080/09243453.2013.871302.

Timmermans, Anneke C, and Sally M Thomas. 2014. "The Impact of Student Composition on Schools' Value-Added Performance: A Comparison of Seven Empirical Studies." *School Effectiveness and School Improvement*. Routledge, 1–12. doi:10.1080/09243453.2014.957328.

Woodhouse, Geoffrey, Min Yang, Harvey Goldstein, and Jon Rasbash. 1996. "Adjusting for Measurement Error in Multilevel Analysis." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 201–12.

Appendix A

Table A1 – Reduced model output from simulation models

	Deterministic Score	Small Error	Medium Error	Large Error
Key Stage 2 Average Point Score (KS2 APS)	17.78	0.92	-9.69	-1.66
KS2 APS Squared	-0.68	0.18	0.81	0.51
KS2 APS Cubed	0.02	0.00	-0.01	-0.01
Constant	116.48	224.27	290.00	247.50
R-squared (Snijders/Bosker - level 1)	1.00	0.93	0.77	0.61

n=535,829 pupils in 3,017 schools

Table A2 – School-level National Pupil Database Variables Used in Section 3

Year	**Value-added Measure**	**KS2 Average Point score**	**Capped GCSE and equivalents point score.**
2004	(VA2NEWE)	(KS2EVAIN)	(PTSCNEWE)
2005-2006	(CVA_KS2)	(CVA2APS)	
2007-2010	(CVA24SCO*SHRINK24)		(TTAPSCP)
2011-2015	(B8VAMEA)	(KS2APS)	
2016	(P8MEA)		(ATT8SCR)

Data for 2011-2016 are readily available to the public from the DfE's performance tables website (DfE, 2015). For 2007-2010, KS4 school-level data were obtained by way of a NPD application. This NPD application also obtained pupil-level data going back to 2004. For 2004-2006, school-level data for these years were obtained by aggregating the 2004-2006 pupil-level data to school-level as mean scores.

Appendix B

To remove the systematic portion of the errors, a follow-on multiple regression of the pupil-level value-added (i.e. the residual unexplained variation) on the true KS2 values was conducted, specified as follows:

3) $$ValueAddedError_{ij} = \beta_0 + \beta_1 KS2APS_{ij} + \beta_2 KS2APS^2_{ij} + \beta_3 KS2APS^3_{ij} + \beta_4 SchoolMean\_KS2APS_{ij} + \beta_5 SchoolMean\_KS2APS^2_{ij} + \varepsilon_{ij}$$

The squared and cubic terms were included to address the non-linearity in the pupil- and school-level error-KS2 relationships. The residual of this regression was saved as an adjusted value-added score with all systematic pupil- and school-level bias removed.

Table B1 – Reduced model output from simulation models

	Small Error	Medium Error	Large Error
Key Stage 2 Average Point Score (KS2 APS)	16.96	27.43	19.01
KS2 APS Squared	-0.86	-1.47	-1.15
KS2 APS Cubed	0.01	0.03	0.02
School Mean KS2 APS	7.89	20.06	22.79
School Mean KS2 APS Squared	-0.16	-0.43	-0.52
Constant	-202.36	-401.03	-360.52
R-squared	0.13	0.32	0.46

n=535,829 pupils in 3,017 schools

The remaining random pupil-level value-added (i.e. the residual) was aggregated to school level as a simple mean school value-added score.